Clustering Vehicles based on Trips Identified from Automatic Number Plate Recognition Camera Scans

Pedro M. Pinto Silva * Matthew Forshaw* A. Stephen McGough*

Abstract

In major cities, government agencies increasingly employ automatic number-plate recognition (ANPR) technology in law enforcement and traffic control. In the Tyne and Wear region (UK) the network of ANPR cameras is used to monitor travel times across sensitive roads. So far, few works have explored the full potential of number-plate scans for analysing individual and collective travel patterns. In this work we present a methodology for deriving trips from vehicle sightings at fixed camera locations. We identify two parameters τ and T as essential for identifying implausible trips and differentiating between multiple trips of the same vehicle. To demonstrate the applicability of trip data we apply *k-means* clustering to trips identified from over 40 million plate scans recorded over fifteen weekdays. Results show that whilst private and transit travel modes can begin to be inferred from the resulting clusters, further work needs to be put into developing a more consistent and integrated framework for trip identification in ANPR data.

1 Introduction

The volume of traffic on our roads has been growing steadily for over 25 years, both in terms of the number of vehicles on the road – increasing by 40.6% in the UK [5] – and the distances covered – 325.5 billion miles driven in the UK in the year ending September 2017 which is up nearly 30% in the last 25 years [6]. This is placing ever more burden on the road infrastructure along with those who police and manage it. In order to better understand how we can deal with this increase in demand we need to better understand how the road network is being used. By understanding road usage we can better deal with congestion, handle traffic incidents, plan road modifications and deal with illegal acts.

In a utopian model we would have full disclosure of all journeys made by all vehicles on the road infrastructure. However, this has numerous ethical and technical issues. From an ethical standpoint should we be allowed to know where all vehicles are at any given point in time? From a technical point of view, although every vehicle could be fitted with a GPS tracker – costly in its own right – there would still exist the issue of how we would collect and stream all of this data for future processing. Alternatively one can view the problem the other way around and rather than tracking individual vehicles look at collecting information by observing vehicles passing points within the road network. A prime example of this approach are Automatic Number Plate Recognition (ANPR) cameras. These cameras are a combination of digital camera coupled with Artificial Intelligence to identify number plates within the image and convert these into strings of characters. ANPR cameras are normally fixed in location[1] able to view all vehicles passing that location.

For ANPR the problem now becomes that of recovering as much information about a vehicle's journey as possible from the limited number of observations. ANPR cameras are normally located on major roads and interchanges, however, this only covers a tiny fraction of the road network. We can, however, estimate routes between cameras by understanding the distances between cameras and the most "sensible" routes between them. This allows us, given a set of ANPR sightings of the same vehicle, to produce a "most likely" route for that journey. It should be noted that we cannot determine the actual start and end of the journey as these will happen in areas not covered by ANPR. It should also be noted that for ethical reasons it is not normal to obtain actual number plates, but rather the hash of these. Though, for most situations this will suffice.

Once we have a set of sightings of a vehicle using ANPR, we now need to convert these into actual journeys. The first requirement is to separate the stream of sightings of a vehicle into individual journeys. Although this can't be done with certainty we can apply general rules to distinguish one journey from the next. For example if two sightings are made from ANPR cameras which are

*School of Computing, Newcastle University, United Kingdom {p.pinto-da-silva2, matthew.forshaw, stephen.mcgough}@newcastle.ac.uk

[1] Although cameras can be in a vehicle and moved from location to location.

connectable by a "sensible" route[2] in a time interval which is "sensible" then these can be determined to be part of the same journey. However, if the timings between two sightings is significantly longer than what would be expected then this would imply that the vehicle stopped between these two cameras and that the later sighting is part of a new journey. The process of journey identification needs to be performed on dirty data which contains numerous impurities which need to be handled. These include:

- **Number plate miss-reads:** Although ANPR cameras have accuracies of around 99%, miss-reads are possible. This can lead to sightings being missed or vehicles being wrongly sighted in locations.
- **Timing errors:** The time-stamps of sightings could be erroneous. The minor side of this is implausible journey times, though, more seriously, this can lead to reordering the set of cameras on a particular journey.
- **Cloned number plates:** For various reasons a number plate may be cloned and used on a different vehicle. This can lead to impossible journeys and journeys that the real vehicle did not make.

Once journeys have been identified from the sightings we can then progress by using these journeys to identify higher-order issues within the road network. In this paper we demonstrate how we can use this journey information in order to identify the most likely class each vehicle is a member of. By clustering over such characteristics as how many journeys are made each day, average length of journeys, the number of different ANPR cameras seen in a day and the times when journeys are made we can cluster vehicles into buses, taxis, commuters and delivery vehicles.

The rest of this paper is presented as follows. In Section 2 we discuss related work. Section 3.1 presents the ANPR data for the Newcastle area. Our process for identifying individual journeys is defined in Section 3.2, while methods for addressing issues resulting from poor camera performance are described in Sections 3.3 and 3.4. Section 3.5 outlines our clustering approach. We report results in Section 4 before offering conclusions and future directions in Section 5.

2 Related Work

The main use of ANPR data for an Urban Traffic Management and Control Centre (UTMC) is estimating average journey times for selected or sensitive links in the road network. Furthermore, several authors, notably Enrique Castillo et al. and Andrew Hazelton et al. have extensively researched how to use number plate data as an extension to link counts for estimating origin-destination matrices and link flows [2, 3, 9]. However, very few works have focused on analysing individual or collective travel patterns from number plate data, particularly across extended periods of time. Moreover, there is no consistent conceptual and analytical framework for transforming number plate data into a historical sequence of trips for each vehicle. Finally, we believe that trip data, properly identified from number plate data, has the potential to unlock a number of new applications for urban traffic control and law enforcement. Thus, in section 3.2 we present a conceptual methodology for grouping multiple camera observations of the same vehicle into one or several trips of that vehicle.

Determining the distribution of travel modes is third of the four fundamental steps in the four-stage model: trip generation, trip distribution, modal split and traffic assignment. The four-stage model is the most widely used traffic modelling methodology for transportation planning [8]. Previous works have used trip information derived from different sources of data to identify travel mode or purpose of trip. More notably, survey data, floating car data and mobile phone data have been used [1, 10]. Although number plate data has been used by Chen et al. (2017) [4] to identify different categories of trips, the authors do not differentiate between private or public travel modes and focus instead on categorising trips by time of occurrence. Hence, in section 3.5 we apply the *k-means* clustering algorithm to derived trip data and based on the results, we discuss the limitations of proposed methods and that of ANPR data.

3 Methodology

3.1 Tyne and Wear ANPR Data Automatic number plate recognition (ANPR) cameras are actively employed in urban traffic environments and play an important role in day-to-day intelligent transportation systems. They can be used by government subsidised entities in urban traffic management and control; by commissioned highway agencies in electronic toll collection; or by law enforcement organisations in detecting speeding vehicles and validating number plate registrations. The wide diversity of applications, paired with the large improvements in price-to-performance ratios of ANPR hardware and software systems, has resulted in increased investments of ANPR cameras for urban environments [7, 12].

[2]Here "sensible" implies that a route between cameras A and B would not need to go through a third camera C.

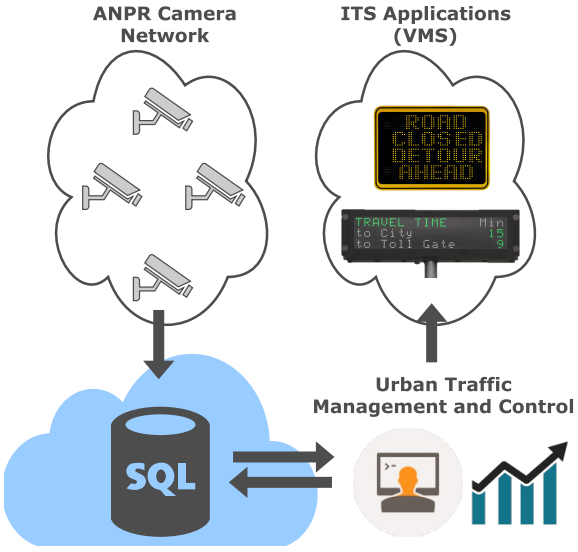


Figure 1: Overview of a ANPR-based system for traffic monitoring and control.

In the region of Tyne and Wear, United Kingdom, there are over 250 active ANPR cameras. Over 1 million license plate detections are recorded by these cameras every day. Figure 2 shows the number of daily scans recorded over a month (February, 2017). Furthermore, every scan is stored in a central database managed by the Tyne and Wear UTMC, and used to compute travel times across particular links of interest in the road network. These are usually major roads that see high volumes of traffic, or road segments more prone to traffic jams. Average journey times can then be conveyed back to the drivers by the way of Variable Message Signs (VMS) or web based applications. Figure 1 represents this interaction.

Number plate data, in its essence, is a stream of events, each representing a vehicle observed by one camera at a specific point in time. An excerpt of the data can be found in Table 1. All number plate were anonymised by the UTMC through a hashing algorithm before the data was shared. Cameras are uniquely identified by an integer and timestamps are relative to each camera's clock. Clock synchronisation is performed using the Network Time Protocol (NTP), as the cameras are connected, through a private network, to a central server. Therefore, the recorded timestamps can be used directly if the synchronisation error is negligible. The following additional information is also captured and provided by each camera: (i) the clock synchronisation error (milliseconds); (ii) the camera's confidence that the identified number plate is the true number plate (percentage); (iii) the direction of travel, away or towards the camera. The confidence in the observation is

Vehicle	Camera	Timestamp	Clock Error	Confidence
169239	1031	1454284800.26	0	100
12862943	18	1454284800.97	8	61
16243894	22	1454284801.46	6	86
4817789	52	1454284803.43	13	94
5503486	110	1454284802.19	22	91
15244177	115	1454284802.83	18	87

Table 1: Sample of number plate data. Clock error is given in milliseconds and confidence as a percentile value.

especially useful as it helps diagnosing license plate recognition errors. On the other hand, the direction of travel is dependent upon the orientation of the camera, which is not provided. Hence, we chose to ignore the latter in this work.

3.2 Trip identification Let the i_{th} sighting of vehicle k be defined as the unordered pair:

$$s_i^k = \{c, t\}, \tag{3.1}$$

where c uniquely identifies a camera, and t is a scalar representing a point in time (e.g. a timestamp).

Let an ordered sequence of sightings of vehicle k define the u_{th} trip of vehicle k:

$$w_u^k = \left(s_{(1)}^k, s_{(2)}^k, \ldots, s_{(n)}^k\right), \tag{3.2}$$

where n is the degree of the trip, i.e. the number of sightings. Moreover, let the corresponding journey time sequence, of degree $n-1$, be defined as the time difference of consecutive sightings:

$$jt_u^k = \left(t_{(2)}^k - t_{(1)}^k, \ldots, t_{(n)}^k - t_{(n-1)}^k\right). \tag{3.3}$$

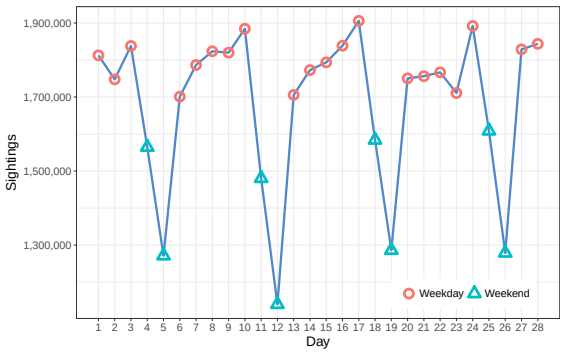
We consider a trip of w_u^k valid under the following conditions:

$$n \geq 1, \tag{3.4}$$

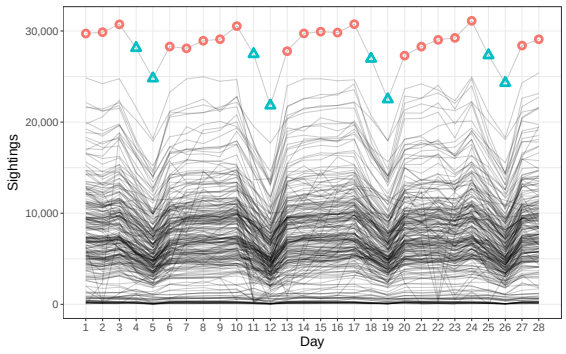
$$\tau < jt_{u(i)}^k < \mathrm{T}\ ,\ \forall i \in \{0, 1, \ldots, n-1\}, \tag{3.5}$$

where τ and T are the lower and upper bound of the i_{th} element of the journey time sequence.

Condition 3.4 is straightforward and specifies that every identified trip should have at least one sighting. Obviously, vehicles can make trips that do not pass through any ANPR cameras and thus have no associated sightings: $n = 0$. However, this work focus on trips that we can observe and hence we consider that $n > 0$.



(a) Total number of scans recorded per day in Tyne and Wear. There is a clear seasonal effect caused by decreasing traffic demands at weekends and increasing traffic volume during weekdays.



(b) Number of scans recorded per ANPR camera and day in Tyne and Wear. Inter-camera variability is observed, as some cameras are located in more traffic intensive road sections than others. Decommissioned or temporarily unavailable cameras (due to loss of power, faulty camera, road closed, etc) can be identified at the bottom.

Figure 2: License plate scans recorded by ANPR cameras during February 2017, in the region of Tyne and Wear, United Kingdom.

Condition 3.5 defines a minimum and maximum travel times between consecutive observations. Its purpose is twofold: (i) first, to allow distinct trips made by the same vehicle to be differentiated. For instance, given two consecutive sightings of k three hours apart, we are likely to want to interpret them as belonging to different trips of k; (ii) second, it allows implausible trips to be identified. For example, an implausible trip can result from observing k at a given camera and then a few seconds later at a second camera, several miles apart. Two explanations are common, either one of the cameras made a detection error, or there is another vehicle with a cloned number-plate travelling on the road network. Evidently, Condition 3.5 is only valid for trips of degree two sightings or greater. Nevertheless, trips can easily be differentiated by first sorting sightings by time of occurrence, then calculating the journey time sequence for the entire sequence and finally comparing each element against T. An example of a trip identified this way can be seen in Figure 4.

The simplest approach to choosing the value of T is to pick a fixed empirical value, such as 5 or 10 minutes. However, if the distance between two cameras is greater than another pair of cameras, then it makes sense that T is relaxed. Similarly, if there is an anomaly in the road network, such as a traffic jam, and the routes connecting the two cameras are affected, then the value of T should also be adapted. Hence, T should be a function of the distance between the two cameras (or, more accurately, of the top n-routes between these) and the distribution of observed journey times. The same rationale can be applied to τ. However, we focus on T and leave τ for analysis at a later stage. Figure 3 shows how the number of trips per degree of trip varies by fixing T at different empirical values.

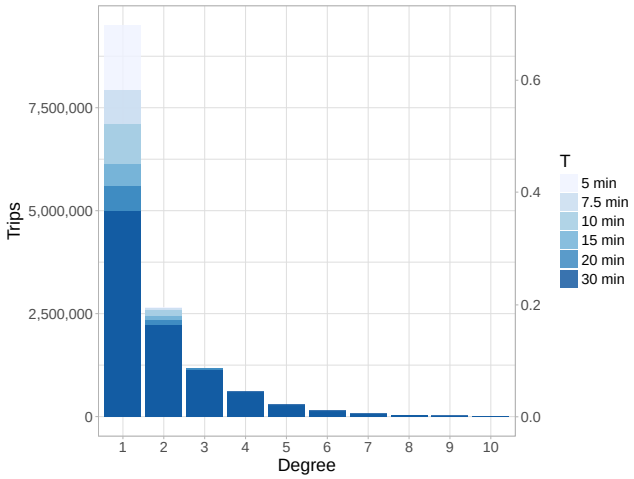


Figure 3: Distribution of trips per degree of trip.

3.3 Duplicate scannings We need to ensure that every trip of vehicle k is unique from all other trips of vehicle k. That is, given W^k the set of all valid trips of k:

$$W^k = \left(w^k_{(1)}, w^k_{(2)}, \ldots, w^k_{(N)}\right), \tag{3.6}$$

where N is the number of trips of k, then there should

Vehicle	Camera	Timestamp	Trip	Sighting	Journey Time	Trip Id
2362920	1014	2017-02-01 00:00:06	1	1	NA	21
2362920	1044	2017-02-01 00:01:28	1	2	82.38	21
2362920	35	2017-02-01 00:02:32	1	3	63.50	21
2362920	32	2017-02-01 00:04:38	1	4	125.95	21

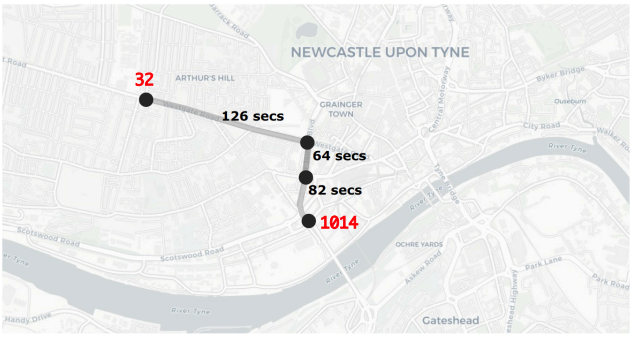


Figure 4: Example of a trip of degree 4. On the left side the corresponding data table is shown. The u_{th} trip of each vehicle is given by the variable *Trip*, whereas the i_{th} sighting is given by variable *Sighting*. The variable *JourneyTime* gives the travel time from the previous to current sighting. Lastly, the variable *TripId* represents the unique sequence of cameras that describes the trip. This allows trips to be grouped and summarised not only in terms of their origins and destination, but also routes. On the right side, the same trip is plotted on a map. The circles represent camera locations whereas the lines represent the fastest driving routes between sightings rather than the true route taken by the vehicle. Even though no routing information is available for each consecutive pair of sightings, the observed journey times can be compared against the distribution of collective journey times to rank the set of most likely routes chosen (which can be determined for instance by Stochastic User Equilibrium [3]).

be no two trips containing the same sighting:

$$s^k_{(u,i)} \neq s^k_{(v,j)}, \forall u, v = 1, 2, \ldots, N \ , \ u \neq v, \tag{3.7}$$

$$\forall i = 1, 2, \ldots, n^u \ , \quad \forall j = 1, 2, \ldots, n^v \tag{3.8}$$

where $s^k_{(u,i)}$ is the i^{th} sighting of the u^{th} trip of vehicle k, and n^u is the degree of u.

In fact, ANPR cameras can identify the same vehicle multiple times in the same pass, if for instance the vehicle is stopped at a junction or traffic light. Hence, if two sightings occurred at the same location in a very short period of time, then there is a strong possibility that these are duplicate observations. As a simplification, we can assume that a trip should not contain cycles and that no camera should appear twice in the same trip. Yet, this assumption ignores cases where a vehicle is required to correct its route by passing through cameras that have already been registered in that trip. Thus, two sightings of vehicle k are different if they were observed: (i) at two different points in time at different locations; (ii) or at the same location with a time interval greater than γ. Otherwise the two sightings are deemed as duplicates:

$$\begin{aligned} &(\ c^k_i \neq c^k_j \ ) \vee \\ &(\ c^k_i = c^k_j \wedge |t^k_i - t^k_j| < \gamma \ ) \ \Rightarrow s^k_i \neq s^k_j \ , \ i \neq j \end{aligned} \tag{3.9}$$

where c^k_i and t^k_i are the camera and timestamp of the i_{th} sighting of vehicle k.

Although the estimation of γ carries similar considerations and consequences as those of estimating T and τ, most duplicates can be identified in consecutive sightings of the same camera within the same trip. Even though a poor estimation of γ also has an impact on error propagation, this decreases substantially after filtering duplicates according to the heuristic above, due to the low occurrence of cycles in trips.

3.4 Errors in plate scanning ANPR cameras have an average accuracy rate of 99.9% or higher. If we consider that on average 1 to 10 out of 10000 number plate scans are misclassified number-plates then approximately between 200 to 2000 scans everyday are incorrect. These errors propagate and lead to innacuracies in the identification of trips. In fact, misclassifications affect the trip sequences of two vehicles: the true passing vehicle and the vehicle erroneously detected instead. The true vehicle will be missing a sighting in the correspoding trip sequence vector whereas the other vehicle's trip sequence will contain an extra invalid sighting. The later may be more easily detected and removed than the former as it may generate a sighting for a vehicle that is normally seen in a different part of the country.

Moreover, ANPR cameras may exceptionally fail to detect passing vehicles. Even though only the passing vehicle is affected in this case, it's trip sequence vector is nonetheless affected. Therefore, due to their impact in trip identification, it is important to detect and

Total Trips	Average Trips	Average Degree	Average Sightings	Average Distinct Origins	Average Distinct Destinations	Average Distinct Routes	Average First Hour	Average Last Hour	Average Hour Difference	Average Rest Time
41	3.42	1.25	5.75	2.75	1.00	3.00	15.33	19.25	3.80	3.70
3	1.50	1.00	1.50	1.50	0.00	1.50	14.00	14.50	0.73	0.73
7	2.33	1.33	3.33	2.33	0.67	2.33	11.00	13.00	2.44	2.41
12	2.40	1.10	3.60	2.40	0.60	2.40	14.40	16.40	2.00	1.95

Table 2: Sample of extracted features from trips taken from 15 weekdays of number plate data.

address missing and misclassified scans. We filter invalid sightings by removing all sightings containing a *confidence* value below 85%. However, we do not address missing scans in this work.

3.5 Clustering vehicles One application of trip data is to group vehicles together according to similar trip patterns, namely frequency and diversity of travel. We can distinguish vehicles based on private and transit modes of transportation and classes within these, such as everyday commuters versus sporadic drivers, buses versus taxis, and other types of vehicles such as lorries and delivery trucks. Ideally, we would like to build a model, through supervised learning, that is able to classify a vehicle from unseen trip data. However, due to privacy laws, there is no publicly available database that maps a plate number into one or several of the categories above. Therefore, we've decided to group vehicles by performing unsupervised learning using the *k-means* clustering algorithm. Despite not being able to perform a rigorous validation or interpretation of the resulting clusters, we can begin to understand how useful trip data is to describe vehicles' travel patterns, given the limited coverage of the cameras in the road network.

Due to distinctly different traffic behaviour during weekends, we only consider trips occurring during weekdays. Therefore, we used all number plate data collected between the 6th and 24th of February and excluded data from the 2 weekends in-between. Furthermore, as mentioned in section 3.2, we used fixed empirical values for T. Table 3 displays the number of trips, average trip degree and the proportion of trips of degree one, for varying values of T. The effect of incrementing T on resulting trips is clear: increased trip degrees and decreased trips containing just a single sighting. On the other hand, we did not set a value for τ, but instead handled implausible trips by filtering out all sightings with confidence below 85%. Duplicates were filtered after identifying consecutive sightings by the same camera occurring within the same trip. Clock synchronisation errors were provided in milliseconds with none exceeding 5 seconds. These were hence ignored.

T	Trips	Average Degree	Proportion of Trips Degree 1
5 min	13,603,759	1.46	0.70
7.5 min	12,394,709	1.60	0.64
10 min	11,690,791	1.69	0.61
15 min	10,823,333	1.83	0.57
20 min	10,305,860	1.92	0.54
30 min	9,653,499	2.04	0.52

Table 3: Overview of trip data for varying values of T.

Transforming trips into features which can be used in clustering algorithms was a 3-step process: (i) First, every trip was summarised as a single row of data. The following information was extracted: degree of trip, origin, destination, route, start and end times. (ii) Second, daily trip information was obtained for each vehicle: number of trips, median of trip degrees, number of sightings, distinct number of origins destinations, and routes, hour of first sighting, hour of last sighting and total rest time between trips. (iii) Finally, daily information per vehicle was collapsed into a single row by averaging this information across the 15 days.

Table 2 depicts a sample of the resulting features vector. A total of 1,034,107 distinct vehicles were detected. However, because there is a high percentage of trips containing a single sighting, some of these features were highly correlated. We therefore, chose to remove three of the features represented in Table 2: *Average Sightings*, *Average Distinct Routes* and *Average Hour Difference*, to avoid the obfuscation of the natural clustering [11]. Furthermore we considered that a trip of degree one has no destination (which explains values of average distinct destination below one) and we filtered all instances of vehicles where the total number of trips is lower than 3, resulting in 642,006 unique vehicles.

Clustering of vehicles was performed using the Hartigan and Wong *k-means* algorithm, for each value of T. The number of clusters k was varied between 2 and 8 and executed with a maximum of 200 iterations and 100 different starting states of the algorithm. The Calinski-Harabasz criterion is used to determine the best value

of k. It minimises and maximises the within-cluster and between-cluster sum of squares respectively, resulting in more natural clusters [11]. The results of trip clustering are presented and discussed in the next section.

4 Results and Discussion

Table 4 provides a summary of multiple runs of *k-means*. For each value of T, the optimal number of clusters is selected by picking the value of k that maximises the Calinski-Harabasz criterion. Furthermore, the measures of inter-cluster (betweeness) and intra-cluster (withiness) are depicted relative to the corresponding total sum of squares (total betweeness and total withiness). It is noteworthy that the best value of k increases inversely to T. Although a higher value of k seems to suggest that trips with smaller values of T are better able to capture the variance of trip data, we have to consider that varying T affects the average trip degree in the same direction whilst affecting the total number of trips in the opposite direction (Table 3).

T	Best k	Betweenss	Average Withinss	Calinski-Harabasz
5 min	8	0.923	0.125	**1,116,962**
7.5 min	7	0.904	0.143	1,003,744
10 min	7	0.896	0.143	911,044
15 min	6	0.865	0.167	794,557
20 min	4	0.786	0.250	754,241
30 min	3	0.710	0.333	743,001

Table 4: *k-means* performance for several values of T.

Cluster	Size	Average Trips	Average Distinct Origins	Average First Hour	Average Last Hour
1	157,962	2.45	2.27	11.38	14.93
2	303,513	2.18	2.09	12.53	14.47
3	21,549	6.02	5.00	8.68	17.23
4	108,094	2.99	2.73	9.99	15.86
5	509	33.62	10.42	5.81	19.67
6	1,993	17.99	11.24	6.45	19.09
7	4,971	10.41	7.64	7.97	18.04
8	58,059	4.09	3.62	9.22	16.60

Table 5: Cluster sizes and centres for T = 5 minutes.

Table 5 depicts the cluster centres for the combination of T and k that maximises the Calinski-Harabasz criterion. These results partially meet our expectations. For instance, we were expecting to find a relatively small cluster representing taxis with a high average number of trips per day, occurring over a variety of origins and destinations and over a large time frame. Clusters 5, 6 and 7 do indeed fit this profile. What differentiates cluster 5 from 6 is the particularly high number of average trips per day. When k is equal to 7 these two clusters merge into one. To some extent, other categories fit this profile as well, namely buses, lorries and delivery trucks. However, we could expect buses to show less diversity in the number of origins and destination as these essentially do multiple runs of the same route throughout the day. Still, this can be explained by the fact that buses take routes through main and secondary roads. As most cameras are placed in main roads, the one long bus trip can be perceived as multiple small trips as the bus alternates between arterial and main roads. This is in fact, one of the downsides of ANPR data and the methodology presented in section 3.2.

On the other hand, we expected to observe a group representing home to work commuters with the first trip of the day starting approximately at eight in the morning and a second trip terminating between five and six in the evening. Although we observe one or two groups with those characteristics, these contain a higher average of trips per day than expected, which may represent for example work to school trips. However, we observe at least one big group of trips occurring mostly during lunch hour. A big contributing factor however is the fact that a high proportion of trips contains only a single sighting (Table 3). It may be the case that many of these drivers choose routes other than those going through ANPR cameras.

To improve the interpretability of these results, we would like to gain access to a governmental database that provides vehicles' weight, wheelplane, make and category from the corresponding plate number. Broadly speaking, vehicle categories are defined relative to the transport of persons or goods, and sub-categories are given according to maximum allowed mass. With this information it would be possible to sort vehicles in the ANPR database into one of the following classes: *Bus, Car, Motorcycle, Truck, Van.* Even though taxis cannot be identified in this way, taxi companies could be contacted directly to obtain this data. However, such databases are available for law enforcement, but not to the public for ethical and privacy reasons. Additionally, this would require us to de-anonymize the hashed number plates. Nevertheless, these results demonstrate that whilst ANPR data is able to capture some of the travel patterns, a clearer and more robust assessment of travel patterns is needed. Furthermore, it is not clear whether this is due to limited coverage of the ANPR cameras in the road network or of the methodology used.

5 Conclusion and Future Work

Most urban cities in the world employ a network of ANPR cameras that are used for law enforcement as well as traffic monitoring and control. Number plate data collected in the Tyne and Wear area is stored and leveraged by the Urban Traffic and Management Centre (UTMC) for computing average journey times across a selection of sensitive roads. However, number plate data could be used more extensively to identify and study individual and collective travel patterns. In this paper, we have presented a set of definitions and constraints that establish a conceptual foundation for identifying vehicle trips from number plate detections.

We have also identified two parameters, τ and T as critical in the discrimination of plausible and implausible trips. Hence, future work should first and foremost focus on developing formal methods to estimate these parameters from observed distributions of travel times and by applying knowledge about the structure of the road network. Any errors in trip identification that occur due to poor estimation and filtering methods will propagate and be amplified in posterior analysis done using trip data. Moreover, methods for addressing issues concerning camera performance, namely wrong and duplicate scans, should be further developed and researched.

Once trip data has been computed, a range of interesting applications are available. This work tries to identify groups of vehicles by clustering information about frequency and diversity of travel. By associating a vehicle with a cluster that represents taxis, or home-to-work commuters, one can begin estimating trip mode usage across the city. However, the results presented here could benefit from extra work and further validation. Namely, gaining access to a database that maps plate numbers to vehicle types would not only provide means to better validate the proposed methodology but also enable the application of supervised learning instead. Furthermore, this work can be improved upon by using other data sources as a baseline for evaluating the reliability of extracted trips, such as GPS taxi traces.

Finally, future work can focus on using trip data to solve interesting research problems such as: (i) real-time route recommendation using probabilistic graphical models; (ii) detection of abnormal trip patterns for helping law enforcement in the identification of suspect vehicles or behaviour; (iii) modelling how drivers make routing choices in the presence of anomalies in the road network.

Acknowledgments The authors would like to thank Phil Blythe, Professor of Intelligent Transport Systems at Newcastle University, and Ray King from the Tyne and Wear Urban Traffic and Management Centre for providing guidance and access to the Automatic Number Plate Recognition database.

References

[1] Alexander, L., Jiang, S., Murga, M., & Gonzlez, M. C. "Origin-destination trips by purpose and time of day inferred from mobile phone data." Transportation research part c: emerging technologies 58 (2015): 240-250.

[2] Castillo, E. Gallego, I., Menndez, J. M., & Rivas, A *Optimal use of plate-scanning resources for route flow estimation in traffic networks*. IEEE Transactions on Intelligent Transportation Systems 11.2 (2010): 380-391.

[3] Castillo, E., Menndez, J. M., & Jimnez, P. *Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations*. Transportation Research Part B: Methodological 42.5 (2008): 455-481.

[4] Chen, H., Yang, C., & Xu, X. *Clustering Vehicle Temporal and Spatial Travel Behavior Using License Plate Recognition Data*. Journal of Advanced Transportation 2017 (2017).

[5] Department for Transport *Provisional Road Traffic Estimates Great Britain: October 2016-September 2017*. url (visited January 2017): https://www.gov.uk-/government/statistics/provisional-road-traffic-estimates-great-britain-october-2016-to-september-2017.

[6] Department for Transport *Vehicle Licensing Statistics: Quarter 3 (Jul-Sep) 2017*. url (visited 31 January 2017): https://www.gov.uk/government/statistics-/vehicle-licensing-statistics-july-to-september-2017.

[7] Hamilton, Andrew, Waterson, B., Cherrett, T., Robinson, A., & Snell, I. *The evolution of urban traffic control: changing policy and technology*. Transportation planning and technology 36.1 (2013): 24-43.

[8] McNally, Michael G. "The four-step model." Handbook of Transport Modelling: 2nd Edition. Emerald Group Publishing Limited, 2007. 35-53.

[9] Parry, K., & Hazelton, M. L. *Estimation of origin-destination matrices from link counts and sporadic routing data*. Transportation Research Part B: Methodological 46.1 (2012): 175-188.

[10] Schssler, N., & Axhausen, K. W. *Identifying trips and activities and their characteristics from GPS raw data without further information*. Arbeitsbericht Verkehrs- und Raumplanung 502 (2008).

[11] Steinley, D., & Brusco, M. J. *Choosing the number of clusters in -means clustering*. Psychological Methods 16.3 (2011): 285.

[12] Zhang, J., Wang, F. Y., Wang, K., Lin, W. H., Xu, X., & Chen, C. *Data-driven intelligent transportation systems: A survey*. IEEE Transactions on Intelligent Transportation Systems 12.4 (2011): 1624-1639.